**Coursera Capstone**

**IBM Applied Data Science Capstone**

Opening a cricket stadium in India

By: Mihir M Kestur



**Introduction**

Cricket is a sport that’s watched and cherished by many in the world. The craze for the sport, in the Indian subcontinent, is arguably unparalleled. The country hosts a variety of international tournaments ranging from the shortest format of the game to the longest 5-day test match series’. It also hosts a number of domestic tournaments, amongst which the Indian premier league (IPL) is a world-class, highly anticipated, annual festival. As per <https://en.wikipedia.org/wiki/Indian_Premier_League>, the Indian cricket industry is easily a billion dollar industry thus having a major positive impact on the economy. Therefore it is easy to realize the importance of having sufficient cricket stadiums to cater to the demand the sport creates in the country.

**Business problem**

The objective of this capstone project is to analyse and select, using data science methodologies and machine learning techniques like clustering, the best location to open a new cricket stadium in the country of India.

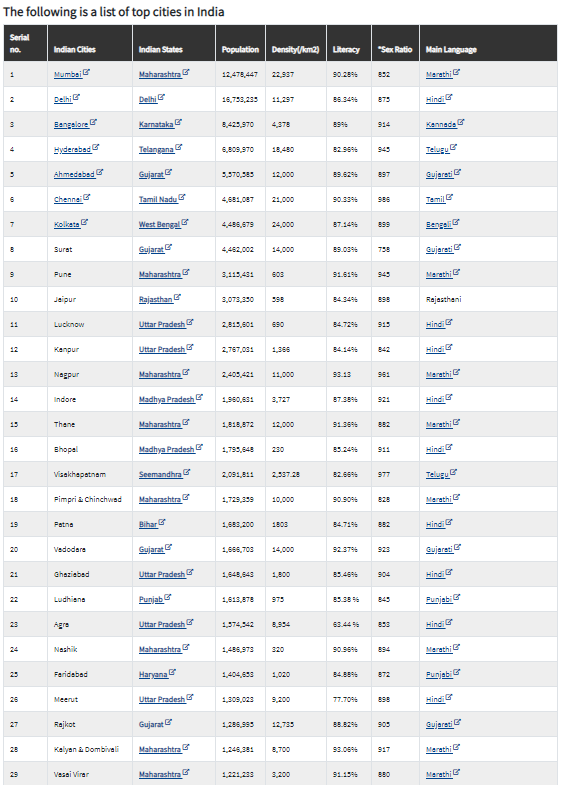
**Data Requirements**

Some parameters that affect the selection of the city are proximity/existence of an airport, restaurants and other attractive venues for the players and spectators. It is also important to exclude the cities in which the cricket stadiums already exist.

* List of top 100 cities in the country web scrapped from <https://www.nriol.com/india-statistics/biggest-cities-india.asp>
* Cities in which cricket stadiums already existing scrapped from <https://en.wikipedia.org/wiki/List_of_international_cricket_grounds_in_India>
* Airport in the city scrapped from

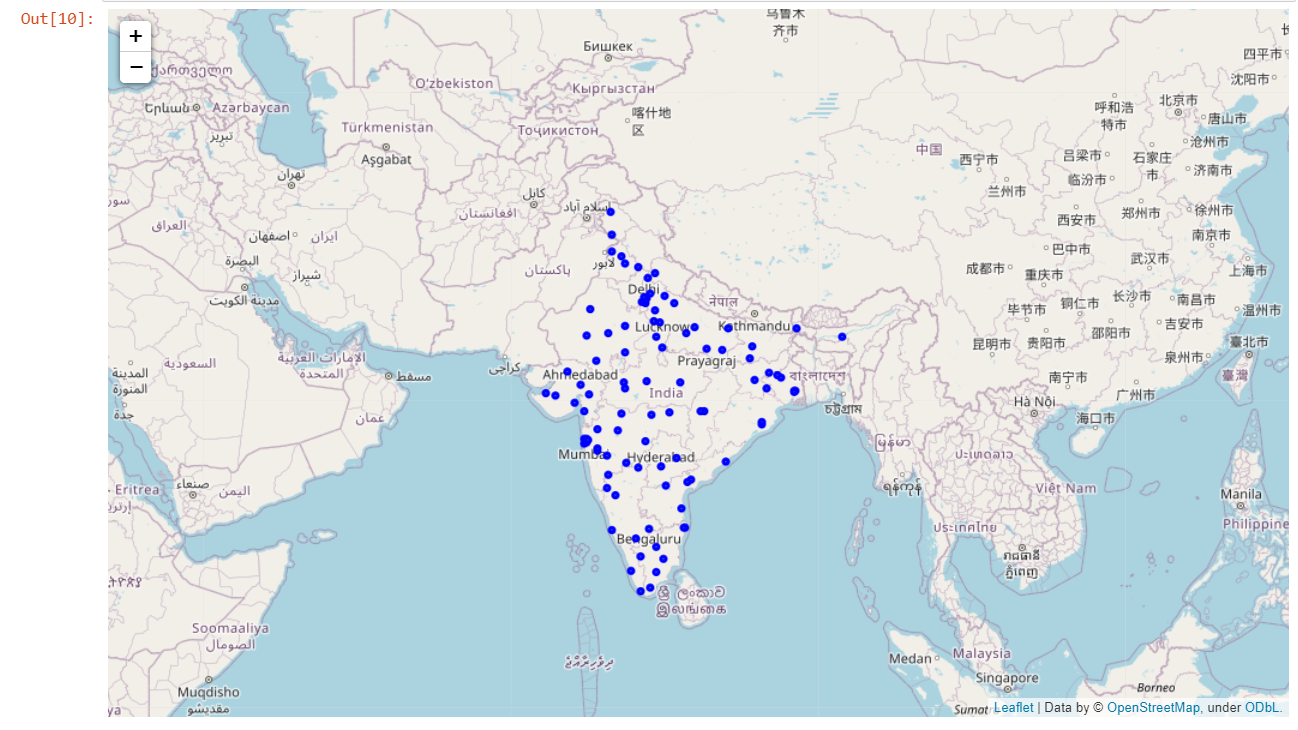
1. <https://www.mapsofindia.com/air-network/international-airport-map.htm>
2. <https://www.mapsofindia.com/air-network/domestic-airport-map.htm>

* Restaurants and other eateries in the city obtained from foursquare API.

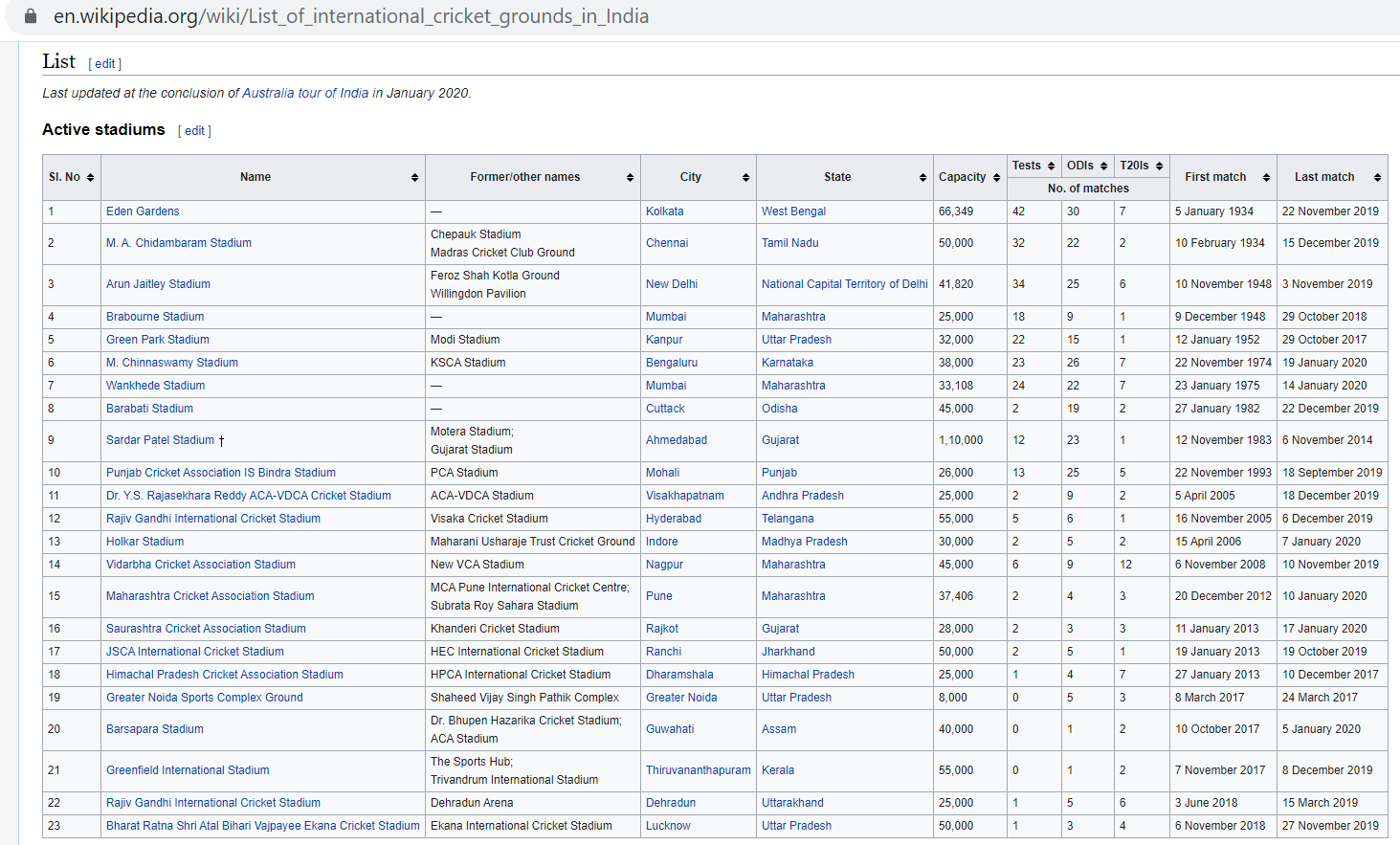
**Methodology**

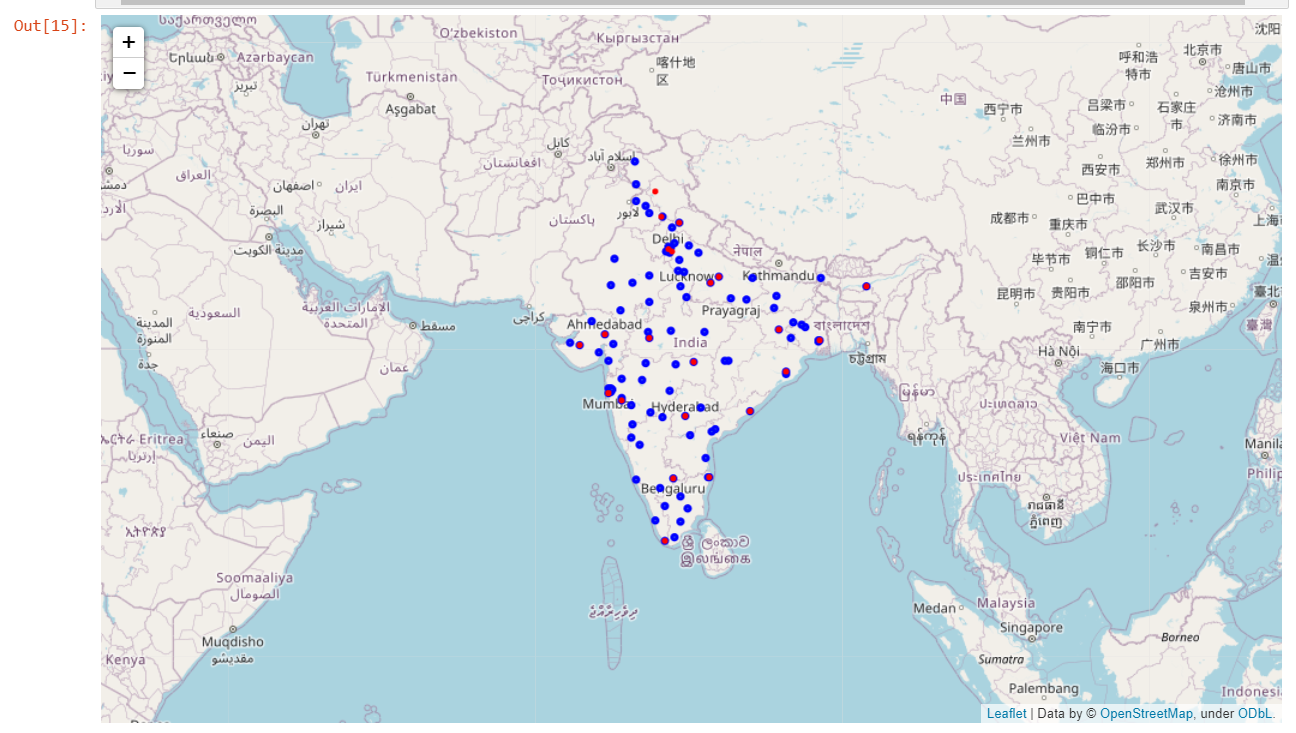
To obtain cities that are good candidates in which a new stadium can be opened, top 100 cities in the country are scrapped from <https://www.nriol.com/india-statistics/biggest-cities-india.asp>. It is done by using the BeautifulSoup library.

It is then plotted on a map to visualize the data. This is done by using the folium library that provides responsive maps with many other features.

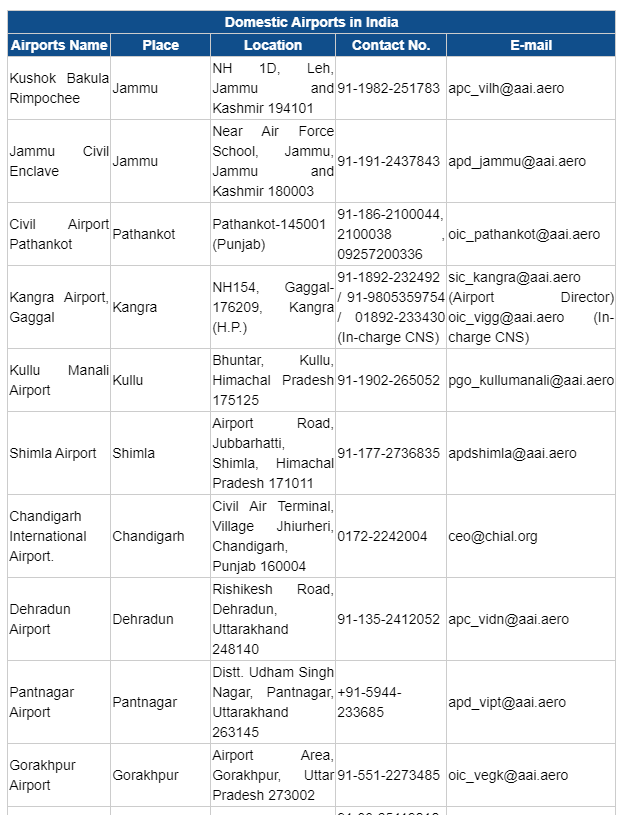
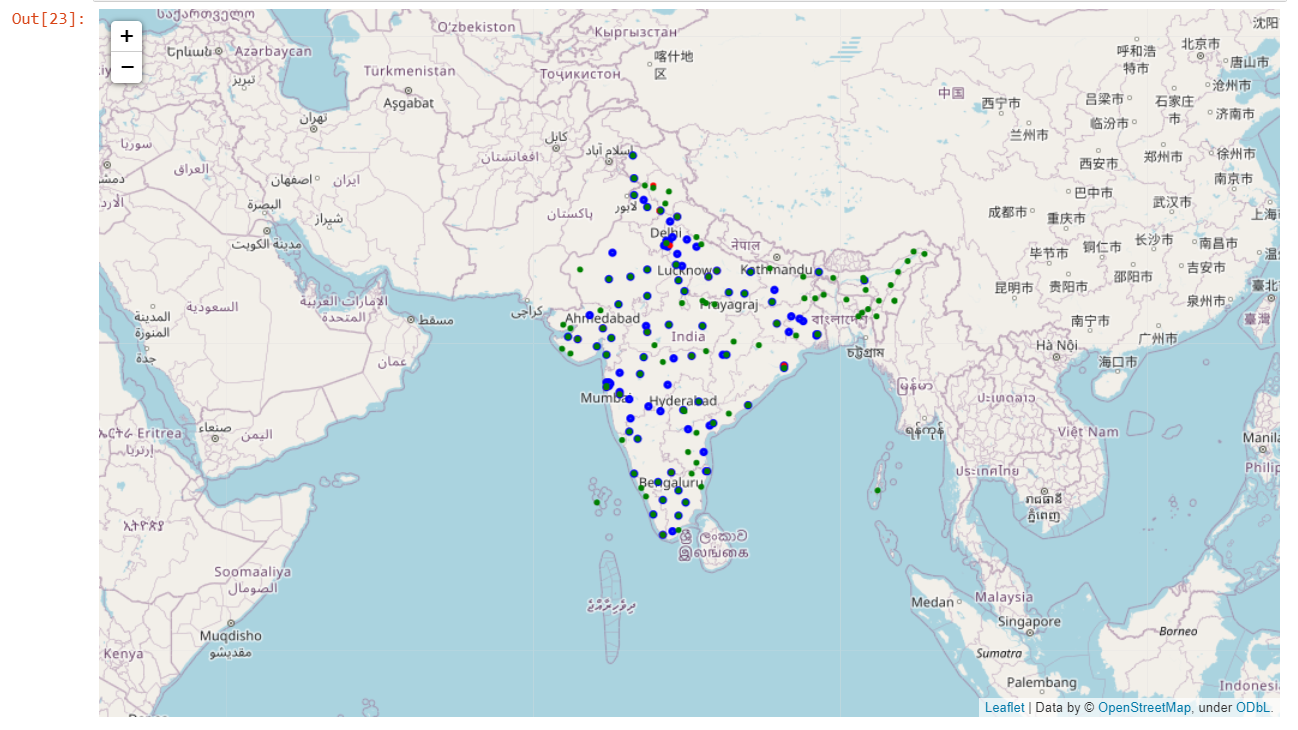


Cities in which cricket stadiums already exist is obtained by web craping a Wikipedia page <https://en.wikipedia.org/wiki/List_of_international_cricket_grounds_in_India>

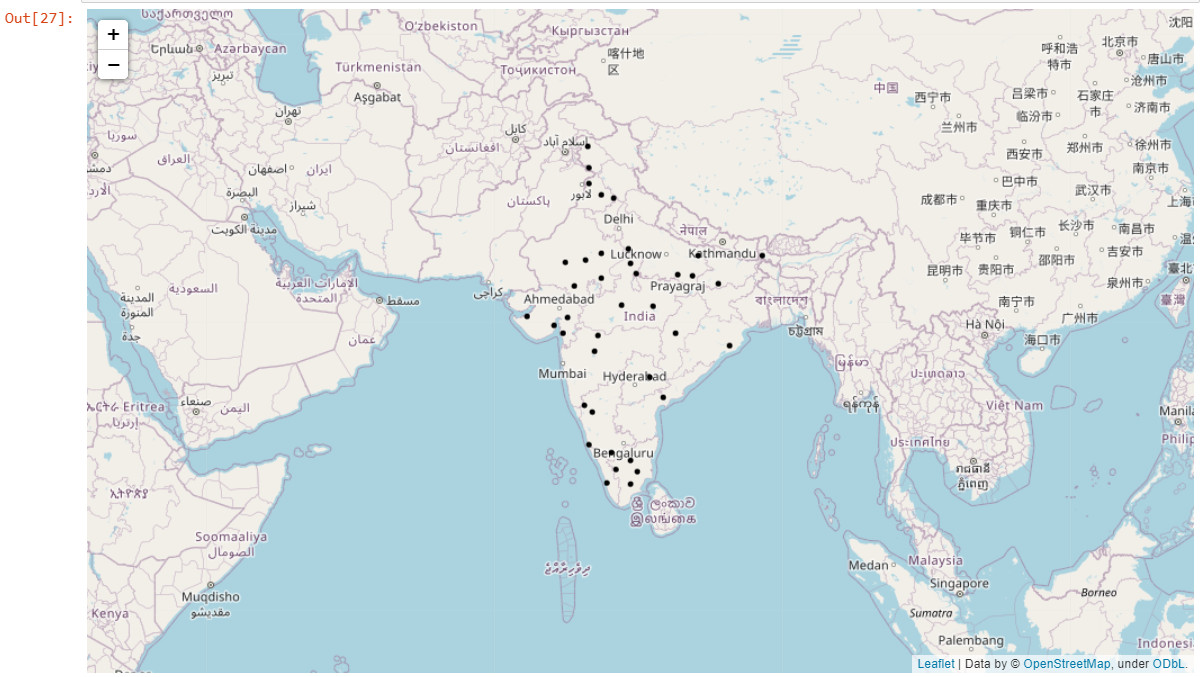
This is done to filter the potential cities and avoid a city to have two fully functional cricket stadiums.

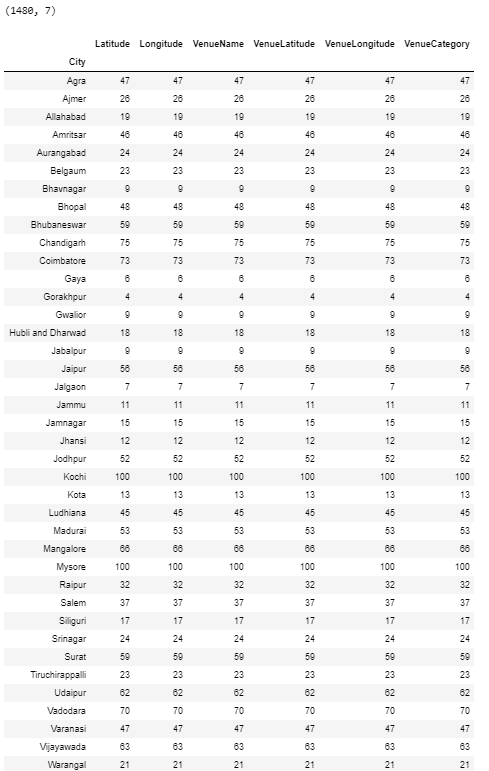
The cities are plotted on the map to visualize. Red points are cities with existing cricket stadiums.

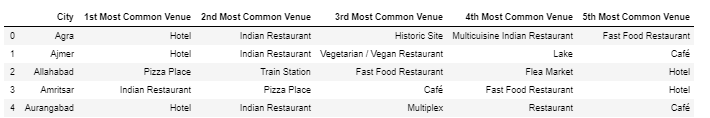
Next, cities which have operational airports are obtained as this an essential infrastructure that allows players, staffs, fans, etc. to move from place to place.

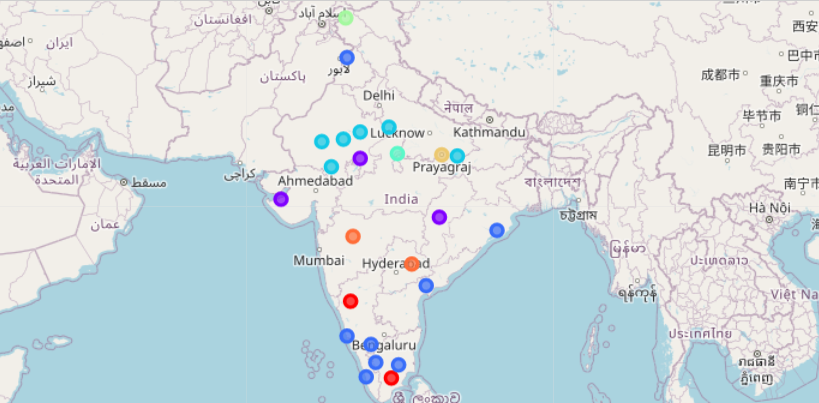
The cities obtained are plotted on the map for visualization and Green markers depict the cities with airports.

This is followed by filtering out cities. Cities which do not have airports and cities in which cricket stadiums exist is removed from the main dataframe that consists of the 100 potential cities, to obtain a filtered dataframe that has 39 potential cities.

It is then plotted on the map for visualization. Black points represent the filtered potential cities.

To further filter the cities, we obtain restuarants and other popular venues using the foursquare API for each city.

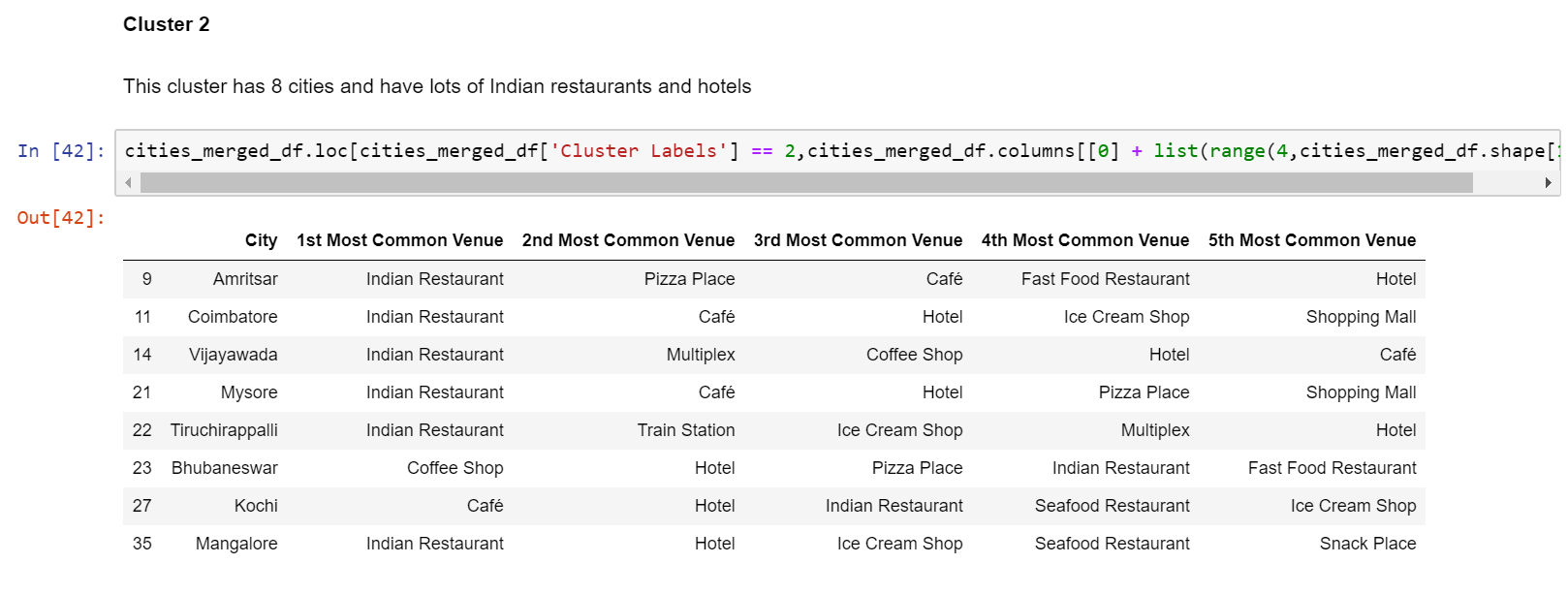
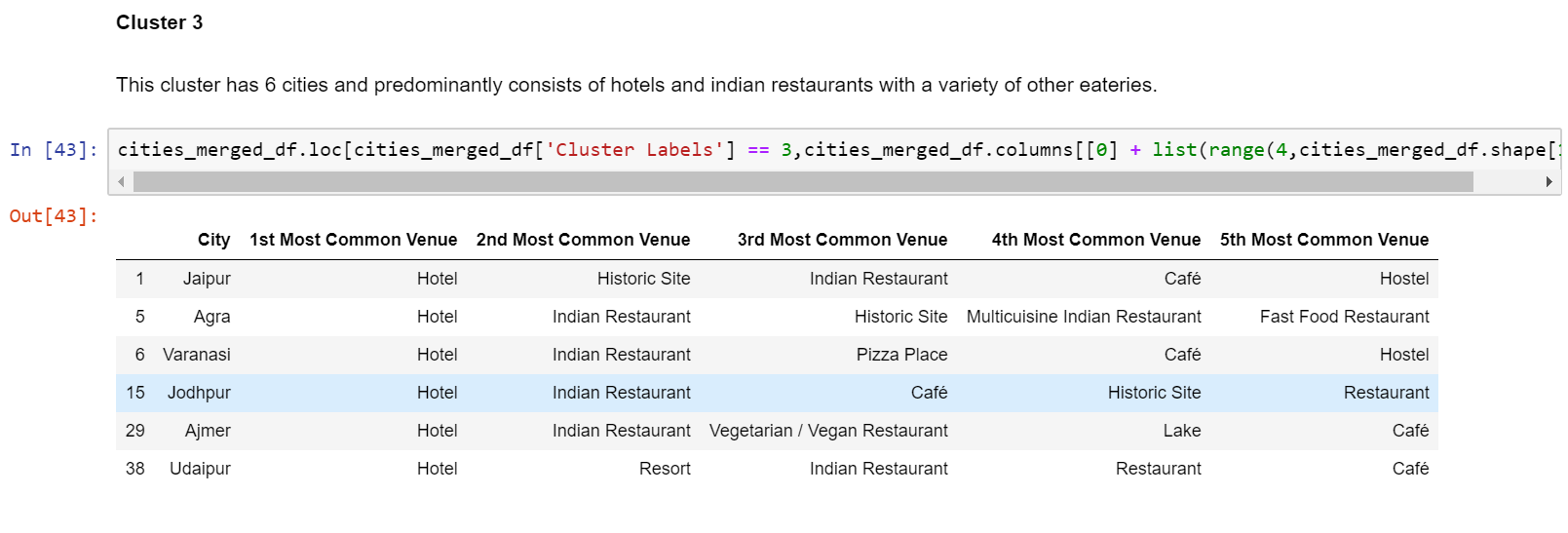
This data is preprocessed via one hot encoding and and top 5 venues in each city is obtained.

The data thus obtained is clusteres via K-means algorithm into 8 different clusters and is visualized on the map.

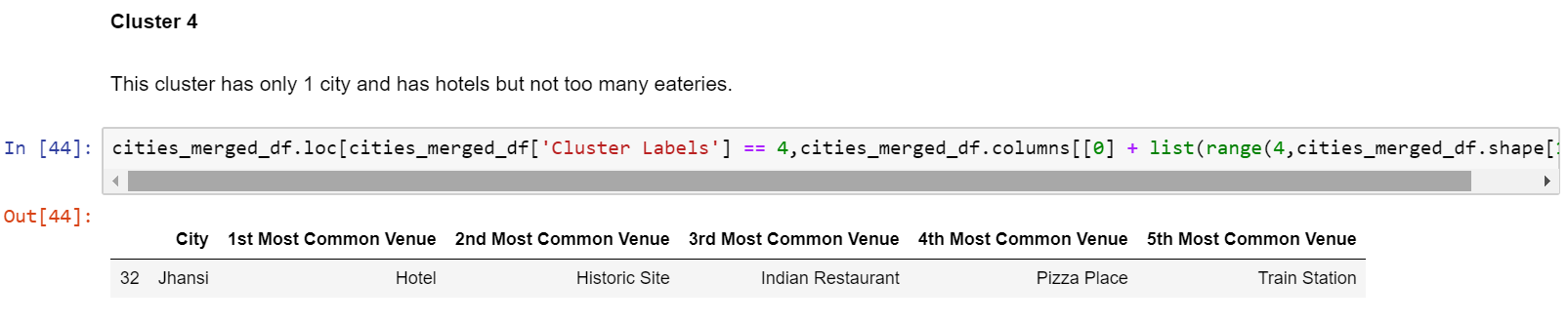
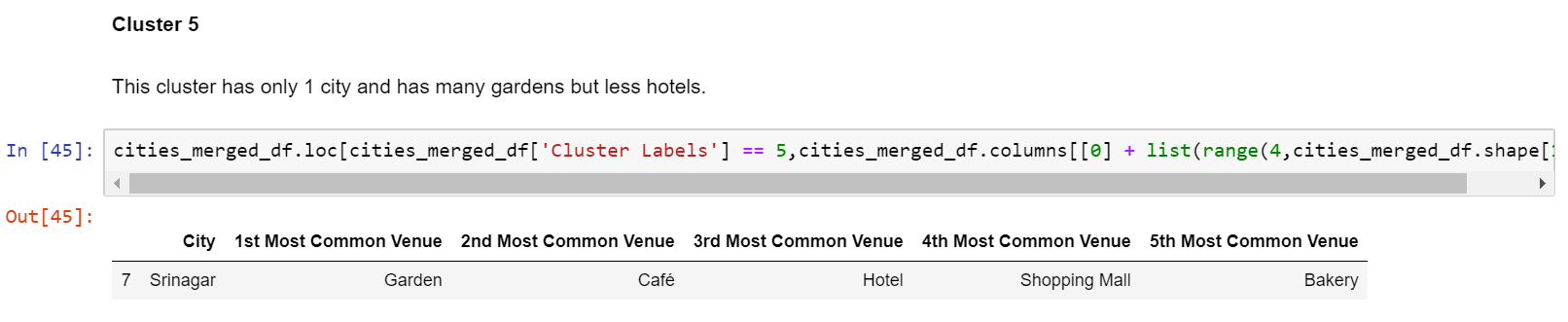
**Observations:**

The obtained clusters is now analyzed individually.

The first cluster has lot of multiplexes and shopping malls, hence it is not ideal to set up a stadium which requires mostly hotels and different eateries.

In the second cluster, cities mainly have Indian resaurants and hotels which makes it a great cluster to look for a potential city.

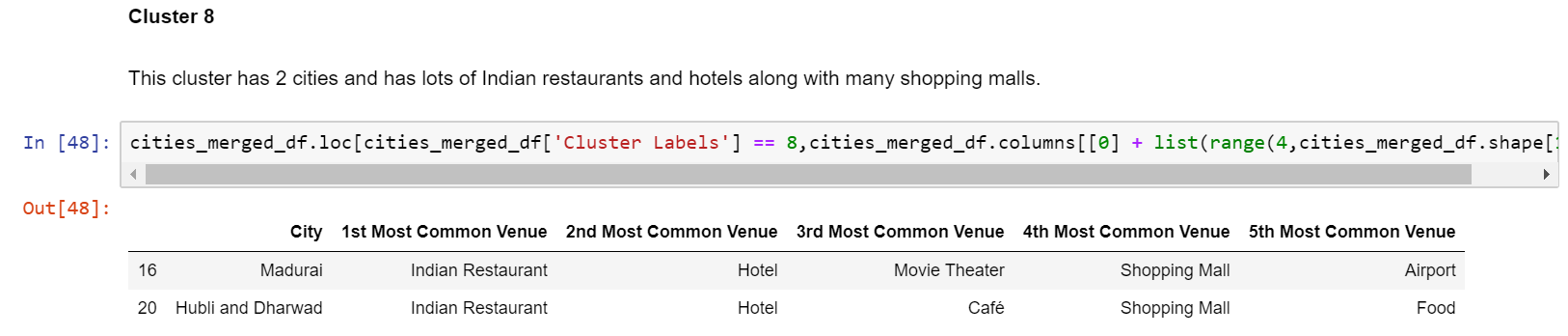
The third cluster has many hotels and a variety of different restaurants. This makes the cluster a strong candidate for building cricket stadiums.

The fourth cluster has one city in which there aren’t many eateries hence this cluster isn’t suitable.

The fifth cluster too has only 1 city in which there are very few hotels and eateries which makes it a non-preferable candidate.

The sixth cluster containing only 1 city has very few hotels (being the 5th most common venue) makes it an undesirable candidate.

The seventh cluster has two cities in which hotels and multiplexes are common. This is a fair candidate and can be considered.

The eighth cluster too has two cities in which Indian restaurants and hotels are common. This makes it a good candidate for the construction of the cricket stadiums.

**Conclusion:**

Thus, we have obtained *8 cities from cluster 2*, *6 cities from cluster 3*, *2 cities from cluster 7*, *2 cities from cluster 8,* that makes 18 potential cities in which a new cricket stadium can be built. We have narraowed down potential candidates from a 100 cities to 18 using clustering algorithm and other exploratory analytic techniques.

*Please note that this analysis is a very primitive and crude form of analysis. Many other parameters like infrastructure, population density, availabilty of technical staff, etc. have not been considered.*